(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－34284
（P2001－34284A）
(43)公開日　平成13年２月９日(2001.2.9)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ１０Ｌ　13/06		Ｇ１０Ｌ　5/04		Ｆ　５Ｄ０４５
13/08		3/00		Ｈ　９Ａ００１

審査請求　未請求　請求項の数11　ＯＬ　（全 12 頁）

(21)出願番号　特願平11－209562

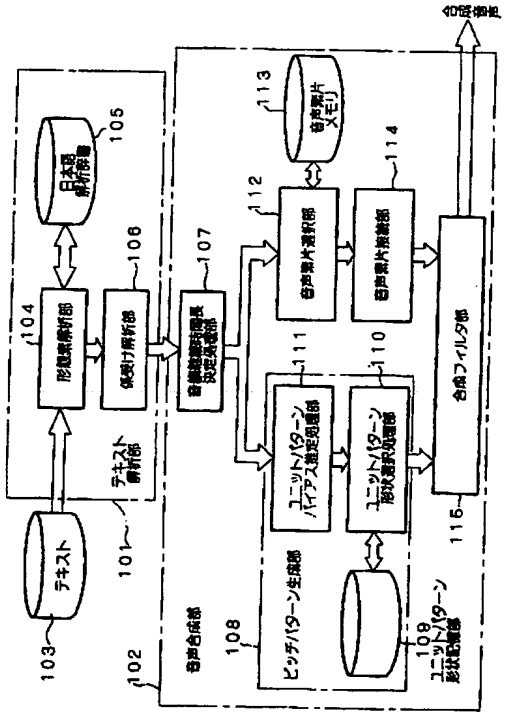
(22)出願日　平成11年７月23日(1999.7.23)

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72)発明者　志賀　芳則
神奈川県川崎市幸区柳町70番地　株式会社東芝柳町工場内
(74)代理人　100058479
弁理士　鈴江　武彦　（外６名）

Ｆターム(参考)　5D045 AA09
9A001 EE05 HH18

(54)【発明の名称】　音声合成方法及び装置、並びに文音声変換プログラムを記録した記録媒体

(57)【要約】
【課題】少数のユニットパターンでありながら、有声音韻と無声音韻の並びによらずに、全ての音韻の並びに適用できるようにする。
【解決手段】人間が発声した音声を分析して得られるピッチパターンを所定の関数に基づいたモデルによって近似し、近似に用いたモデルパラメータより生成される近似されたピッチパターンから、所定の単位で切り出したユニットパターンのピッチ周波数最大値を基準としたユニットパターン形状を記憶部１０９に複数記憶し、所定の規則を用いて入力テキストに対するテキスト解析部１０１の解析結果に基づき、選択処理部１１０にて記憶部１０９からユニットパターン形状を選択すると共に、推定処理部１１１にて所定の規則を用いてピッチバイアス量を推定し、このユニットパターン形状とピッチバイアス量からピッチパターン生成部１０８でユニットパターンを生成し、そこからピッチパターンを生成する。



1

【特許請求の範囲】

【請求項１】　入力テキストデータを解析してその解析結果に基づいてピッチパターンを生成し、生成したピッチパターンに基づいて音声を合成する音声合成方法において、

人間が発声した音声を分析して得られるピッチパターンを所定の関数に基づいたモデルによって近似し、近似に用いたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから、所定の単位で切り出したユニットパターンをそのまま、もしくはクラスタリングによって得られる代表ユニットパターンの形で複数記憶しておき、

前記入力テキストデータの解析結果に基づいて、前記記憶しておいた複数のユニットパターンの中から使用するユニットパターンを前記所定の単位毎に選択し、この所定の単位毎に選択したユニットパターンに基づいてピッチパターンを生成することを特徴とする音声合成方法。

【請求項２】　入力テキストデータを解析してその解析結果に基づいてピッチパターンを生成し、生成したピッチパターンに基づいて音声を合成する音声合成方法において、

人間が発声した音声を分析して得られるピッチパターンを所定の関数に基づいたモデルによって近似し、近似に用いたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから、所定の単位で複数のユニットパターンを切り出し、この切り出した複数のユニットパターンをその形状と形状のピッチバイアス値に分離することで、前記人間が発声した音声の内容を表わすテキストデータを解析した結果に基づいて、ユニットパターンの形状を選択するための第１の規則とユニットパターンの形状のピッチバイアス値を推定するための第２の規則とを予め作成しておく一方、

前記近似されたピッチパターンから所定の単位で切り出したユニットパターンの形状をそのまま、もしくはクラスタリングによって得られる代表ユニットパターン形状の形で複数記憶しておき、

前記第１の規則を用いることにより、前記入力テキストデータの解析結果に基づいて、使用するユニットパターンの形状を前記記憶しておいた複数のユニットパターン形状の中から前記所定の単位毎に選択すると共に、前記第２の規則を用いることにより、前記入力テキストデータの解析結果に基づいて、使用するユニットパターン形状のピッチバイアス値を前記所定の単位毎に推定し、

前記選択したユニットパターン形状と前記推定したピッチバイアス値とに基づいて前記所定の単位毎にユニットパターンを生成し、この所定の単位毎に生成したユニットパターンに基づいてピッチパターンを生成することを特徴とする音声合成方法。

【請求項３】　前記近似されたピッチパターンから所定の単位で切り出したユニットパターンのピッチ周波数最

2

大値を、対応する前記ユニットパターン形状のピッチバイアス値とすることを特徴とする請求項２記載の音声合成方法。

【請求項４】　前記入力テキストデータの解析結果に基づくユニットパターンの選択では、合成すべきモーラ数またはアクセント型に最も近いと判定される発声がされた人間の音声のピッチパターンを前記モデルで近似したピッチパターンから切り出したユニットパターンを選択し、この選択したユニットパターンをそのままもしくは変形して用いることを特徴とする請求項１記載の音声合成方法。

【請求項５】　前記入力テキストデータの解析結果に基づくユニットパターン形状の選択では、合成すべきモーラ数またはアクセント型に最も近いと判定される発声がされた人間の音声のピッチパターンを前記モデルで近似したピッチパターンから切り出されたユニットパターンの形状を選択し、この選択したユニットパターン形状をそのままもしくは変形して用いることを特徴とする請求項２記載の音声合成方法。

【請求項６】　前記入力テキストデータの解析結果に基づくユニットパターンの選択に際し、前記所定の単位毎にユニットパターンの候補を複数選択し、この選択した所定の単位毎のユニットパターン候補の全ての組み合わせについて、前記入力テキストデータを合成するための接続部分の歪みに対して所定の評価を行って、この評価が最も良好となるユニットパターン候補の組み合わせを選択し、この選択した組み合わせの各ユニットパターンを接続してピッチパターンを生成することを特徴とする請求項１記載の音声合成方法。

【請求項７】　前記入力テキストデータの解析結果に基づくユニットパターン形状の選択に際し、前記所定の単位毎にユニットパターン形状の候補を複数選択すると共にピッチバイアス値の候補を複数推定し、

前記選択した複数のユニットパターン形状候補及び推定した複数のピッチバイアス候補を個別もしくは同時に、前記入力テキストデータを合成するための接続に対して所定の評価を行って、この評価が最も良好となるユニットパターン形状とピッチバイアス値を前記所定の単位毎に１つ決定し、

前記所定の単位毎に決定したユニットパターン形状とピッチバイアス値に基づいて前記所定の単位毎にユニットパターンを生成し、この前記所定の単位毎に生成したユニットパターンに基づいてピッチパターンを生成することを特徴とする請求項２記載の音声合成方法。

【請求項８】　入力テキストデータを解析してテキスト解析結果を生成するテキスト解析手段と、

人間が発声した音声を分析して得られるピッチパターンを近似する所定の関数に基づいたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから、所定の単位で切り出されたユニッ

トパターンがそのまま、もしくはクラスタリングによって得られる代表ユニットパターンの形で複数記憶されたユニットパターン記憶手段と、
前記テキスト解析手段のテキスト解析結果に基づいて前記ユニットパターン記憶手段からユニットパターンを前記所定の単位毎に選択し、この所定の単位毎に選択したユニットパターンに基づいてピッチパターンを生成するピッチパターン生成手段と、
前記ピッチパターン生成手段によって生成されたピッチパターンに基づいて音声を合成する音声合成手段とを具備することを特徴とする音声合成装置。

【請求項９】　入力テキストデータを解析してテキスト解析結果を生成するテキスト解析手段と、
人間が発声した音声を分析して得られるピッチパターンを近似する所定の関数に基づいたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから、所定の単位で切り出されたユニットパターンの形状がそのまま、もしくはクラスタリングによって得られる代表ユニットパターン形状の形で複数記憶されたユニットパターン形状記憶手段と、
前記入力テキストデータの解析結果に基づいてピッチパターンを生成するピッチパターン生成手段と、
前記ピッチパターン生成手段によって生成されたピッチパターンに基づいて音声を合成する音声合成手段とを具備し、
前記ピッチパターン生成手段は、
人間が発声した音声を分析して得られるピッチパターンを近似する所定の関数に基づいたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから、所定の単位で複数のユニットパターンを切り出して、この複数のユニットパターンをその形状と形状のピッチバイアス値とに分離することで、前記人間が発声した音声の内容を表わすテキストデータを解析した結果に基づいて予め作成された、ユニットパターンの形状を選択するための第１の規則及びユニットパターンの形状のピッチバイアス値を推定するための第２の規則のうちの前記第１の規則を用いることにより、前記入力テキストデータの解析結果に基づいて、使用するユニットパターンの形状を前記記憶手段から前記所定の単位毎に選択するユニットパターン形状選択手段と、
前記第２の規則を用いることにより、前記入力テキストデータの解析結果に基づいて、使用するユニットパターンの形状のピッチバイアス値を前記所定の単位毎に推定するピッチバイアス値推定手段とを備え、
前記ユニットパターン形状選択手段により選択されたユニットパターンの形状と前記ピッチバイアス値推定手段により推定されたピッチバイアス値に基づいて前記所定の単位毎にユニットパターンを生成し、この所定の単位毎に生成したユニットパターンに基づいてピッチパターンを生成することを特徴とする音声合成装置。

【請求項１０】　コンピュータに、
入力テキストデータを解析するテキスト解析ステップと、
前記テキスト解析ステップでの解析結果に基づいてピッチパターンを生成するピッチパターン生成ステップであって、人間が発声した音声を分析して得られるピッチパターンを近似する所定の関数に基づいたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから、所定の単位で切り出されたユニットパターンがそのまま、もしくはクラスタリングによって得られる代表ユニットパターンの形で複数記憶された記憶手段から、前記入力テキストデータの解析結果に基づいて、使用するユニットパターンを前記所定の単位毎に選択する選択ステップと、
前記選択ステップで前記所定の単位毎に選択したユニットパターンに基づいてピッチパターンを生成するピッチパターン生成ステップとを実行させるための文音声変換プログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項１１】　コンピュータに、
入力テキストデータを解析するテキスト解析ステップと、
人間が発声した音声を分析して得られるピッチパターンを近似する所定の関数に基づいたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから、所定の単位で複数のユニットパターンを切り出して、この複数のユニットパターンをその形状と形状のピッチバイアス値とに分離することで、前記人間が発声した音声の内容を表わすテキストデータを解析した結果に基づいて予め作成された、ユニットパターンの形状を選択するための第１の規則及びユニットパターンの形状のピッチバイアス値を推定するための第２の規則のうちの前記第１の規則を用いることにより、前記所定の単位で切り出されたユニットパターンの形状がそのまま、もしくはクラスタリングによって得られる代表ユニットパターン形状の形で複数記憶された記憶手段から、前記テキスト解析ステップでの解析結果に基づいて、使用するユニットパターンの形状を前記所定の単位で選択するユニットパターン形状選択ステップと、
前記第２の規則を用いることにより、前記テキスト解析ステップでの解析結果に基づいて、使用するユニットパターンの形状のピッチバイアス値を前記所定の単位で推定するピッチバイアス値推定ステップと、
前記ユニットパターン形状選択ステップで選択したユニットパターンの形状と前記ピッチバイアス値推定ステップで推定したピッチバイアス値に基づいて前記所定の単位毎にユニットパターンを生成し、この所定の単位毎に生成したユニットパターンに基づいてピッチパターンを生成するピッチパターン生成ステップとを実行させるための文音声変換プログラムを記録したコンピュータ読み

5

取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、入力テキストデータの解析結果から、合成すべき音声の韻律を生成するのに好適な音声合成方法及び装置、並びに文音声変換プログラムを記録した記録媒体に関する。

【０００２】

【従来の技術】この種の音声合成装置の代表的なものに、音声を細分化して蓄積し、その組み合わせによって任意の音声を合成可能な規則合成装置があることが知られている。以下では、規則合成装置の従来技術の例を図を参照しながら説明していく。

【０００３】図５は従来の規則合成装置の構成を示すブロック図である。図５の規則合成装置（音声合成装置）は、入力されるテキストデータ（以下、単にテキストと称する）を音韻情報と韻律情報からなる記号列に変換し、その記号列から音声を生成する文音声変換（Text-to-speech conversion ：以下、ＴＴＳと称する）処理を行う。

【０００４】この図５の規則合成装置におけるＴＴＳ処理機構は、大きく分けて言語処理部１２と音声合成部１３の２つ処理部からなり、日本語の規則合成を例に取ると次のように行われるのが一般的である。

【０００５】まず言語処理部１２では、テキストファイル１１から入力されるテキスト（漢字かな混じり文）に対して形態素解析・構文解析等の言語処理を行い、形態素への分解、係受け関係の推定等の処理を行うと同時に、各形態素に読みとアクセント型を与える。その後言語処理部１２では、アクセントに関しては複合語等のアクセント移動規則を用いて、読み上げの際のアクセントの区切りとなる句（以下、アクセント句と称する）毎のアクセント型を決定する。

【０００６】次に音声合成部１３内では、得られた「読み」に含まれる各音韻の継続時間長を音韻継続時間長決定処理部１４にて決定する。音韻継続時間長は、日本語特有の拍の等時性に基づき決定する手法が一般的である。例えば、子音の継続時間長は子音の種類により一定とし、各モーラの基準時刻である子音から母音へのわたり部の間隔が一定になるように、母音の継続時間長が決定される。

【０００７】続いて、上記のようにして得られる「読み」に従って、音韻パラメータ生成処理部１６が音声素片メモリ１５から必要な音声素片を読み出し、読み出した音声素片を上記の方法で決定した音韻継続時間長に従って、時間軸方向に伸縮させながら接続して、合成すべき音声の特徴パラメータ系列を生成する。

【０００８】ここで、音声素片メモリ１５には、予め作成された多数の音声素片が格納されている。音声素片は、アナウンサ等が発声した音声を分析して、スペクトルの包絡特性を表現する所定の音声の特徴パラメータを得た後、所定の合成単位、本従来例では日本語の音節の単位（子音＋母音：以下、ＣＶと称する）で、日本語の音声に含まれる全ての音節を上記特徴パラメータから切り出すことにより作成される。また本従来例では、前記の特徴パラメータとしてケプストラムの低次の係数を利用している。低次のケプストラム係数は次のようにして求めることができる。まず、アナウンサ等が発声した音声データを、一定幅・一定周期で窓関数（例えばハニング窓）で切り出し、各窓内の音声波形に対してフーリエ変換を行い音声の短時間スペクトルを計算する。次に、得られた短時間スペクトルのパワーを対数化して対数パワースペクトルを得たのち、対数パワースペクトルを逆フーリエ変換する。こうして計算されるのがケプストラム係数である。ここで、ケプストラムの特性として、高次の係数は音声の基本周波数情報を、低次の係数は音声のスペクトル包絡情報を保持していることはよく知られている。

6

【０００９】音声合成部１３では更に、ピッチパターン生成処理部１７が上記アクセント型をもとに、図６に示すように、ピッチの高低変化が生じる音節時間長の中心時刻に点ピッチを設定する。こうして、設定された複数の点ピッチ間を直線補間して、アクセント句毎でピッチのアクセント成分を生成する。更に、人間の発話時のピッチの自然下降を表現するピッチのイントネーション成分を生成し、これに前記生成したピッチのアクセント成分を重畳してピッチパターンを生成する。

【００１０】最後に、合成フィルタ処理部１８において、有声区間ではピッチパターンに基づいた周期パルスを、無声区間ではホワイトノイズをそれぞれ音源とし、音声の特徴パラメータ系列から算出したフィルタ係数を用いてフィルタリングを行い所望の音声を合成する。ここでは、合成フィルタ処理部１８の合成フィルタとして、ケプストラム係数を直接フィルタ係数とするＬＭＡ（Log Magnitude Approximation）フィルタ（対数振幅近似フィルタ）を用いている。

【００１１】

【発明が解決しようとする課題】上記した規則合成装置に代表される従来の音声合成装置では、その音声合成装置で生成される韻律に関して次のような問題があった。

【００１２】まず、ピッチパターン生成処理部１７でのピッチパターンの生成において、１音節あたり１点ピッチというかなり粗い近似を行っているため、合成音声のピッチ変化が、人間の発声した音声のそれとかなり違ったものとなってしまい、合成音声の自然性が損なわれていた。

【００１３】この問題に対して、例えば特開平６－２３６１９７号公報、及び特開平９－３４４９２号公報では、人間の発声した音声のピッチパターンからアクセント句単位でユニットパターンを切り出して記憶してお

き、音声合成時にそのユニットパターンを検索・配置して、滑らかなピッチパターンを得ようとする方法が提案されている。

【００１４】ところが、これら方法にも次のような問題がある。まず、人間の発声した音声のピッチパターンは、無声子音等においてピッチの存在しない区間（無声区間）がある。そこで特開平９－３４４９２号公報では、アクセント句内において、有声音韻と無声音韻の並びが一致するユニットパターンを利用することを提案している。しかしこれでは、例えば、ある有声音韻と無声音韻の並びのユニットパターンが１つしかない場合、同じ並びの音声を合成しようとすると、この１つのパターンが絶えず選択されることになり、バリエーションが乏しく、そのために不自然なピッチが生じる虞がある（第１の問題）。これを解決するためには、アクセント句内の様々な位置に無声区間をもつユニットパターンを用意すればよいが、これでは多くのユニットパターンが必要となり、それらを格納するメモリが膨大になる。

【００１５】また、特開平９－３４４９２号公報では、該当する有声音韻と無声音韻の並びがユニットパターン群の中に存在しない場合は、無声区間のピッチを前後の有声区間のピッチから補間し利用する手法が提案されている。しかしながら、無声区間とピッチの存在する有声区間の境界時点は、人間の発声における声帯の開放と閉鎖（による振動）の境界時点となるため、ピッチに乱れが生じており、そのまま補間するとピッチに不自然なゆれが生じてしまう。このユニットパターンの無声区間と有声区間の境界におけるピッチの乱れが、音声合成時のピッチ生成において悪影響を及ぼす問題（第２の問題）は、特開平６－２３６１９７号公報における提案手法でも起こり得る問題であり、これが合成音声のピッチを不自然にする原因となっていた。

【００１６】また特開平９－３４４９２号公報には、音声合成しようとするアクセント句と、モーラ数やアクセント型が同じである自然発声された音声のピッチをユニットパターンとする方法が記載されている。この方法では、あらゆるモーラ数とアクセント型の組み合わせを網羅するユニットパターンを用意することから、それらを格納するために多くのメモリが必要となっていた（第３の問題）。

【００１７】また、特開平９－３４４９２号公報と特開平６－２３６１９７号公報においては、合成しようとする各アクセント句に対してどのユニットパターンを使用するか、或いはユニットパターンをどのように制御するかは、隣接アクセント句に使用されるユニットパターンとその制御に無関係に決定するため、隣接アクセント句間のピッチが不連続になる虞があった（第４の問題）。

【００１８】このように、ユニットパターンを利用した従来の音声合成手法には種々の欠点があった。

【００１９】本発明は上記事情を考慮してなされたものでその目的は、少数のユニットパターンを用意するだけで、有声音韻と無声音韻の並びによらずに、全ての音韻の並びに適用でき、様々な文を対象とする音声合成が、ユニットパターンの無声区間と有声区間の境界におけるピッチの乱れに起因した音声合成時の揺れやピッチ制御の悪影響を生じることなく行える音声合成方法及び装置、並びに文音声変換プログラムを記録した記録媒体を提供することにある。

【００２０】本発明の他の目的は、少数のユニットパターンでありながら、様々な合成内容に対応可能な音声合成方法及び装置、並びに文音声変換プログラムを記録した記録媒体を提供することにある。

【００２１】

【課題を解決するための手段】本発明の第１の特徴は、入力テキストデータを解析してその解析結果に基づいてピッチパターンを生成し、生成したピッチパターンに基づいて音声を合成する音声合成方法において、人間が発声した音声を分析して得られるピッチパターンを所定の関数に基づいたモデルによって近似し、近似に用いたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから、所定の単位で切り出したユニットパターンをそのまま、もしくはクラスタリングによって得られる代表ユニットパターンの形で複数記憶しておき、（文音声変換の対象となる）入力テキストデータの解析結果に基づいて、上記記憶しておいた複数のユニットパターンの中から使用するユニットパターンを所定の単位毎に選択し、この所定の単位毎に選択したユニットパターンに基づいてピッチパターンを生成するようにしたことにある。

【００２２】このように本発明においては、人間が発声した音声を分析して得られるピッチパターンを所定の関数に基づいたモデルによって近似し、近似に用いたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから切り出されるユニットパターンを利用することで、少数のユニットパターンを用意するだけで、有声音韻と無声音韻の並びによらずに、全ての音韻の並びに適用でき、ピッチ分析の不具合を回避して、無声区間における補間を確実に行うことが可能となる。つまり少数のユニットパターンを用意するだけで、様々な文を対象とする音声合成が、ユニットパターンの無声区間と有声区間の境界におけるピッチの乱れに起因した音声合成時の揺れやピッチ制御の悪影響を生じることなく行える。

【００２３】ここで、上記近似に用いたモデルパラメータ、もしくは上記近似されたピッチパターンから切り出される複数のユニットパターンをそのまま用いる構成（第１の構成）では、無声区間における補間をより確実に行うことができる。これに対し、複数のユニットパターンをそのまま用いずに、クラスタリングによって得られる代表ユニットパターンを用いる構成（第２の構成）

では、この効果は上記第１の構成ほどではないものの、近似されたピッチパターンから切り出されるユニットパターンの特徴を極力保持しながら、ユニットパターンを記憶するのに必要な記憶容量を減らすことができる。

【００２４】本発明の第２の特徴は、人間が発声した音声を分析して得られるピッチパターンを所定の関数に基づいたモデルによって近似し、近似に用いたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから、所定の単位で複数のユニットパターンを切り出し、この切り出した複数のユニットパターンをその形状と形状のピッチバイアス値に分離することで、上記人間が発声した音声の内容を表わすテキストデータを解析した結果に基づいて、ユニットパターンの形状を選択するための第１の規則とユニットパターンの形状のピッチバイアス値を推定するための第２の規則とを予め作成しておく一方、上記所定の単位で切り出したユニットパターンの形状をそのまま、もしくはクラスタリングによって得られる代表ユニットパターン形状の形で複数記憶しておき、上記第１の規則を用いることにより、入力テキストデータの解析結果に基づいて、使用するユニットパターンの形状を上記記憶しておいた複数のユニットパターン形状から所定の単位毎に選択すると共に、上記第２の規則を用いることにより、入力テキストデータの解析結果に基づいて、使用するユニットパターン形状のピッチバイアス値を所定の単位毎に推定し、選択したユニットパターン形状と推定したピッチバイアス値とに基づいて所定の単位毎にユニットパターンを生成し、この所定の単位毎に生成したユニットパターンに基づいてピッチパターンを生成することにある。ここで、上記近似されたピッチパターンから所定の単位で切り出したユニットパターンのピッチ周波数最大値を、対応するユニットパターン形状のピッチバイアス値とするとよい。

【００２５】このように本発明においては、ユニットパターンの形状とバイアス値を分離することによって、ピッチ分析の不具合を回避して、無声区間における補間が確実に行えるだけでなく、様々な合成内容に対応可能となる。

【００２６】ここで、上記モデル、即ち人間が発声した音声を分析して得られるピッチパターンを近似するモデルには、臨界制動２次線形系の応答関数に基づいたピッチ生成モデル、或いはスプライン関数に基づいたモデルが適している。

【００２７】本発明の第３の特徴は、入力テキストデータの解析結果に基づくユニットパターン（またはユニットパターンの形状の）の選択では、合成すべきモーラ数またはアクセント型に最も近いと判定される発声がされた人間の音声のピッチパターンを上記モデルで近似したピッチパターンから切り出したユニットパターン（またはユニットパターンの形状）を選択するようにして、この選択したユニットパターン（ユニットパターン形状）をそのままもしくは変形して用いるようにしたことにある。

【００２８】本発明においては、ピッチ分析の不具合を回避し、無声区間における補間が確実に行えるようになるだけでなく、合成すべきモーラ数、アクセント型と全く同じユニットパターンが存在しなくてもピッチパターンの生成が可能となる。つまり、あらゆるモーラ数とアクセント型のユニットパターンを用意することなく、比較的少数のユニットパターンで、全てのモーラ数・アクセント型を網羅し音声を合成することが可能となる。特に、選択したユニットパターン（ユニットパターン形状）を変形して用いる構成とした場合、無声区間における補間がより確実に行える。

【００２９】本発明の第４の特徴は、上記入力テキストデータの解析結果に基づくユニットパターンの選択に際し、所定の単位毎にユニットパターンの候補を複数選択し、この選択した所定の単位毎のユニットパターン候補の全ての組み合わせについて、上記入力テキストデータを合成するための接続部分の歪みに対して所定の評価を行って、この評価が最も良好となるユニットパターン候補の組み合わせを選択し、この選択した組み合わせの各ユニットパターンを接続してピッチパターンを生成するようにしたことにある。

【００３０】本発明においては、接続歪みを考慮して、接続歪みが最も少ないユニットパターン候補の組み合わせを用いることで、より滑らかな接続のピッチパターンを得ることが可能となる。つまり、ユニットパターン間において、不自然なピッチの歪みを生じさせず、滑らかにピッチを繋ぎ、聞きやすい合成音声を提供できる。

【００３１】本発明の第５の特徴は、入力テキストデータの解析結果に基づくユニットパターン形状の選択に際し、所定の単位毎にユニットパターン形状の候補を複数選択すると共にピッチバイアス値の候補を複数推定し、この選択した複数のユニットパターン形状候補及び推定した複数のピッチバイアス候補を個別もしくは同時に、上記入力テキストデータを合成するための接続に対して所定の評価を行って、この評価が最も良好となるユニットパターン形状とピッチバイアス値を所定の単位毎に１つ決定し、この所定の単位毎に決定したユニットパターン形状とピッチバイアス値に基づいてユニットパターンを生成し、この生成したユニットパターンに基づいてピッチパターンを生成することにある。

【００３２】本発明においては、ユニットパターンの形状とバイアス値を分離するだけでなく、接続歪みを考慮して、接続歪みが最も少ないユニットパターン形状とピッチバイアス値との候補の組み合わせを用いることで、様々な文の音声合成に対し、常に滑らかな接続のピッチパターンを得ることが可能となる。

【００３３】以上により本発明においては、聞き取りや

すく、長時間聞いていても疲れない、自然な抑揚をもつ音声を合成することが可能となる。

【００３４】なお、以上の方法に係る本発明は装置（音声合成装置）に係る発明としても成立する。また、本発明は、コンピュータに当該発明に相当する手順を実行させるための（或いはコンピュータを当該発明に相当する手段として機能させるための、或いはコンピュータに当該発明に相当する機能を実現させるための）文音声変換プログラムを記録した記録媒体（コンピュータ読み取り可能な記録媒体）としても成立する。

【００３５】

【発明の実施の形態】以下、本発明の実施の形態につき図面を参照して説明する。図１は本発明の一実施形態に係る音声の規則合成装置の概略構成を示すブロック図である。

【００３６】図１の音声規則合成装置（以下、音声合成装置と称する）は、例えば、パーソナルコンピュータ等の情報処理装置上で、ＣＤ－ＲＯＭ、フロッピーディスク、メモリーカード等の記録媒体、或いはネットワーク等の通信媒体により供給される専用のソフトウェア（文音声変換ソフトウェア）を実行することにより実現されるもので、テキストから音声を生成する文音声変換（ＴＴＳ）処理機能を有しており、その機能構成は、大別してテキスト解析部１０１と、音声合成部１０２とに分けられる。

【００３７】テキスト解析部１０１は、文音声変換の対象となる入力文である漢字かな混じり文を解析して語の同定を行い（形態素解析）、得られた品詞情報等をもとに、文の構造を推定し（係受け解析）、その結果を出力する処理を司る。一方、音声合成部１０２は、テキスト解析部１０１の出力であるテキスト解析結果をもとに音声を生成する処理を司る。

【００３８】さて、図１の音声合成装置において、文音声変換（読み上げ）の対象となるテキスト（ここでは日本語文書）はテキストファイル１０３として保存されている。本装置では、文音声変換ソフトウェアに従い、当該ファイル１０３から漢字かな混じり文を読み出して、テキスト解析部１０１及び音声合成部１０２により以下に述べる文音声変換処理を行い、音声を合成する。

【００３９】まず、テキストファイル１０３から読み出された漢字仮名混じり文（入力文）は、テキスト解析部１０１内の形態素解析部１０４に入力される。形態素解析部１０４は、入力される漢字かな混じり文に対して形態素解析を行ない、読み情報とアクセント情報を生成する。形態素解析とは、与えられた文の中で、どの文字列が語句を構成しているか、そしてその語の文法的な属性がどのようなものかを解析する作業である。

【００４０】形態素解析部１０４は、入力文を日本語解析辞書１０５と照合して全ての形態素系列候補を求め、その中から、文法的に接続可能な組み合わせを出力する。ここで日本語解析辞書１０５には、形態素解析時に用いられる情報（例えば文法情報）と共に、文の最小構成要素である個々の「形態素」の読みとアクセント型が登録されている。そのため、形態素解析により形態素が定まれば、同時に読みとアクセント型を与えることができる。

【００４１】次に、テキスト解析部１０１内の形態素解析部１０４にて決定した文に含まれる個々の語の文法属性は、同じテキスト解析部１０１内の係受け解析部１０６に出力され、各語の係受け関係を推定する文構造の解析が行われる。

【００４２】以上のようにして、テキスト解析部１０１では、語の読みやアクセントの情報と共に品詞や係受け関係の情報を出力する。これらの情報は、音声合成部１０２に渡される。

【００４３】音声合成部１０２では、まず音韻継続時間長決定処理部１０７が起動される。この音韻継続時間長決定処理部１０７での音韻継続時間長決定には、［従来の技術］の欄で述べた図５中の音韻継続時間長決定処理部１４と同様に、日本語特有の拍の等時性に基づき決定する手法を採用している。このため、音韻継続時間長決定処理部１０７の処理内容の説明は省略する。

【００４４】音声合成部１０２内の音韻継続時間長計算処理部１０７により入力文（入力テキスト）の読みに含まれる各音韻の継続時間長、更に詳細に述べるならば各モーラの（子音部並びに母音部の）継続時間長が決定されると、同じ音声合成部１０２内のピッチパターン生成処理部１０８が起動される。

【００４５】ピッチパターン生成処理部１０８は、テキスト解析部１０１により決定されたアクセント情報、品詞、係受け情報に基づいて、合成すべき音声のピッチパターンを生成する。そのためピッチパターン生成処理部１０８は、予め用意されたユニットパターンの形状を複数格納したユニットパターン形状記憶部１０９がアクセス可能なようになっており、テキスト解析部１０１からの入力に基づいて、使用するユニットパターンの形状をユニットパターン形状記憶部１０９から選択するユニットパターン形状選択処理部１１０と、選択したユニットパターンの形状からピッチパターンを生成する際に、ユニットパターンに加算するピッチバイアス量を推定するユニットパターンバイアス推定処理部１１１とを有する。

【００４６】ここで、一旦音声合成の処理の説明から離れ、予め用意しておかなければならないユニットパターン形状の作成方法と、これを選択するための規則の構築方法と、ピッチバイアス量の推定方法について説明する。

【００４７】まず初めに、所定のテキストを人間（例えばアナウンサ）が読み上げた音声を収録する。収録した音声をコンピュータに取り込んで分析し、声の高さのパ

ターンを表すピッチ（または基本周波数）パターンを抽出する。ピッチパターンの抽出には、自己相関を用いる方法、ケプストラムを用いる方法など様々な手法が提案されており、それらのうちの適当な手法を利用すればよい。

【００４８】こうして得られたピッチパターンのうち、無声子音等無声区間の存在する箇所では、通常図２（ａ）に示されるように、有声区間と無声区間の境界付近に細かな変化が存在する。

【００４９】図２（ａ）に示すようなピッチ変化は、ピッチの合成単位（ユニットパターン）中で無声区間位置が一致するような音声を合成しようとするときには問題とならないかもしれない。しかし、無声区間の位置が一致しないときは、ピッチのない無声区間のピッチを適当に決めなければならない。このとき、無声区間におけるピッチを単に有声区間端のピッチから補間して求めると、図２（ａ）に破線２１，２２，２３で示したように、有声区間と無声区間の境界のピッチ変動の影響でピッチが細かく上下してしまい、合成された音声に震えが生じてしまう。

【００５０】そこで本実施形態では、平滑化の関数に基づいたモデルで近似することで、この有声区間と無声区間の境界付近の細かなピッチ変化を取り除くようにしている。ピッチパターンのモデルとしては、音響学会誌（1971年）、Vol.27, p445–453に記載された藤崎らによる臨界制動２次線形系が良く知られており、また、音響学会講演論文集、平成10年9月，p217–218に示された森川らによるスプライン関数を用いたものも有効である。本実施形態では、前者の臨界制動２次線形系のモデルを用いてピッチパターンを近似する。

【００５１】前記論文記載の手法に基づき、モデルにより近似されたピッチパターンは、図２（ｂ）に示されるように滑らかで、無声区間のピッチも得られるため、無声区間を補間する必要がなくなる。

【００５２】こうして得られた近似されたピッチパターンから、図２（ｃ）に示されるように、ユニットパターンとしてアクセント句単位で切り出す。そして、切り出したユニットパターンの最大周波数値（最大ピッチ周波数値）をユニットパターン形状のピッチバイアス量（対数値）とし、最大周波数値をピッチ周波数の基準にしたパターン相対値をユニットパターン形状とする（図２（ｃ））。ここでは、図２（ｃ）に示されるように、１モーラあたり４点のピッチとして時間長の正規化を行う。

【００５３】次に、収録時に用意したテキストをテキスト解析して得られる結果と、ユニットパターン形状、ピッチバイアス量との対応関係を規則化する。前記した収録内容に十分な分量があれば、統計的な手法を用いることで、自動的にユニットパターン形状、ピッチバイアス量を推定する規則を構築できる。

【００５４】本実施形態では、ピッチバイアス量については、テキスト解析結果である文を構成するアクセント句のアクセント型、モーラ数、品詞、係受け関係の各情報、つまりピッチバイアス量への影響が大きい情報を説明変数とし、ピッチバイアス量を外的基準として、「数量化Ｉ類」を適用し、ピッチバイアス量を予測する規則を自動的に構築する。一方、ユニットパターン形状については、品詞情報と係受け情報、更にピッチバイアス量（実測値）を説明変数とし、「回帰木」の手法を用いて、最適なユニットパターンを選択する規則（回帰木）を自動構築する。

【００５５】これらの統計手法は一般的に良く知られており、「数量化Ｉ類」については、「数量化理論とデータ処理」林知己夫他著、朝倉書店（１９８２）などに、「回帰木」については、“Classification and Regression Trees”, L. Breiman他著、Wadsworth Statistics/Probability Series, USA(1984)などに代表される書物に理論や実践方法が記載されているため、ここでは説明を省略する。

【００５６】但し、ユニットパターン形状を推定するユニットパターン形状選択規則を作成する際、アクセント型とモーラ数により個別のユニットパターン形状選択規則を作成すると共に、最終的に分割された説明変数空間に複数のユニットパターン形状のデータが含まれるように、分割は適当な場所で止められる。

【００５７】上述した最終的に分割された個々の説明変数空間に含まれる複数のユニットパターン形状は、類似性が高い。そこで、これらをクラスタリングし、同一クラスに属する（分類される）ものを平均化して１つのユニットパターン形状で代表させることで、つまりユニットパターンの個数を減らすことで、メモリ容量の低減と処理の簡略化を図ることもできる。しかし、ここでは後に説明する処理のために、平均化することは避け、分割された個々の説明変数空間に複数のユニットパターン形状を持たせておく。

【００５８】ピッチパターン生成処理部１０８では、以上のようにして構築されたユニットパターン形状のピッチバイアス量推定規則、及びユニットパターン形状選択規則を用い、テキスト解析部１０１から渡されるアクセント情報、品詞情報、係受け情報から、合成すべき音声のピッチパターンを次のように生成する。

【００５９】まず初めに、数量化Ｉ類を適用して得られたユニットパターン形状のピッチバイアス量推定規則をピッチパターン生成処理部１０８内のユニットパターンバイアス推定処理部１１１にて用いて、アクセント情報、品詞情報、係受け情報からユニットパターン形状のピッチバイアス量を推定する。

【００６０】次に、回帰木を適用して得られたユニットパターン形状選択規則をピッチパターン生成処理部１０８内のユニットパターン形状選択処理部１１０にて用い

て、ユニットパターン形状記憶部１０９からユニットパターン形状を選択する。この、ユニットパターン形状の選択に当たっては、まず合成しようとするアクセント句のモーラ数とアクセント型から推定に利用するユニットパターン形状選択規則を決定し、テキスト解析部１０１の出力である品詞情報、係受け情報に併せて、既に推定されたピッチバイアス量を用い、ユニットパターン形状選択規則に基づいてユニットパターン形状を選択する。

【００６１】ここで、もし、最初の利用するユニットパターン形状選択規則を決定する際に、モーラ数とアクセント型が一致するユニットパターン用の形状選択規則が存在しない場合は、「ユニットパターン形状を選択しようとしているアクセント句のモーラ数以上で最も近いモーラ数のユニットパターン用の形状選択規則のうち、アクセント型が最も近いユニットパターン用の形状選択規則」を用いる。

【００６２】但し、アクセント型の異なるユニットパターン用の形状選択規則を用いるため、規則によって選択されたユニットパターン形状はアクセント型が異なっている。そこで、図３に示すように、アクセント型が一致するような変形を加える。この図３の例は、モーラ数が７モーラ、アクセント型が５型のユニットパターン用の形状選択規則として、モーラ数が６モーラ、アクセント型が４型のユニットパターン用の形状選択規則が選択された場合を示したもので、当該選択規則で選択された６モーラ４型のユニットパターンの形状が、７モーラ５型に変形されている。

【００６３】このようにすることで、ユニットパターン形状選択規則構築時の収録音声データの中に、完全にモーラ数とアクセント型が合致するアクセント句が存在しなくても、ピッチパターンを生成することが可能になる。

【００６４】上述したように、ユニットパターン形状選択規則のための回帰木において、分割された説明変数空間には複数のデータが含まれる。このためユニットパターン形状選択処理部１１０での選択の対象となるユニットパターン形状として、アクセント句毎に複数の候補が挙げられる。

【００６５】そこでユニットパターン形状選択処理部１１０では、アクセント句毎に、複数の形状候補の中からユニットパターンを１つ選び、図４に示されるように、推定されたピッチバイアス量を加算したうえで、文のピッチパターンを得るべくユニットパターンの接続を試す。このときユニットパターン形状選択処理部１１０は、ユニットパターン接合部分のピッチ周波数の差（歪み）ｄｉを計測し、これを２乗して、文内の全ての接合部の２乗歪みｄｉ^{2}の総和Σｄｉ^{2}を計算する。そしてユニットパターン形状選択処理部１１０は、ユニットパターン形状候補の全ての組み合わせに対してこの計算を行い、最も２乗歪みの総和が小さい組み合わせを選択する。こうすることで、最も接続の滑らかな文ピッチパターンを生成するユニットパターン形状の組み合わせを決めることができる。

【００６６】このようにして、ユニットパターン形状の組み合わせを決定すると、その決定した組み合わせに、先に接続を試したときと同様に、アクセント句毎に既に決定しているピッチバイアス量を加算し、接続することで、合成すべき音声のピッチパターンを生成することができる。この接続の際、歪みを極力減らすためピッチの平滑処理を行ってもよい。

【００６７】なお、ピッチバイアス量についてもアクセント句毎に複数の候補を推定するようにして、全ての組み合わせについて、上記のユニットパターンの組み合わせと同様の手法で、文全体で（つまり合成するための接続に対して）歪みの評価を行って、評価が最も良好なピッチバイアス量の組み合わせ選択するようにしてもよい。この場合、ユニットパターン形状とピッチバイアス量について個別に評価するのではなく、各アクセント句毎に得られる、ユニットパターンの候補と、ピッチバイアス量の候補の全ての組み合わせのユニットパターンを生成し、つまり候補とした選択されたユニットパターンに候補として選択されたピッチバイアス量を加算してユニットパターンの候補を生成する処理を全ての組み合わせについて行い、それを全アクセント句分について行って、各アクセント句毎にユニットパターン候補を求め、それを各アクセント句間で接続して上記の評価を行うことで、各アクセント句毎に、最適なユニットパターン形状とピッチバイアス量を同時に決定することも可能である。

【００６８】一方、音声合成部１０２内の音声素片選択部１１２は、テキスト解析部１０１から渡されるアクセント句毎の読みに基づく音声素片選択を行う。本実施形態では、サンプリング周波数１１０２５Ｈｚで標本化した実音声を改良ケプストラム法により窓長２０msec、フレーム周期１０msecで分析して得た０次から２５次の低次ケプストラム係数を、子音＋母音（ＣＶ）の単位で、日本語音声の合成に必要な全音節を切り出した計１３７個の音声素片が蓄積された音声素片ファイル（図示せず）が用意されている。この音声素片ファイルの内容は、文音声変換ソフトウェアに従う文音声変換処理の開始時に、例えばメインメモリ（図示せず）に確保された音声素片領域（以下音声素片メモリと称する）１１３に読み込まれる。

【００６９】そこで音声素片選択部１１２は、音声素片の選択を音声素片メモリ１１３を対象に行う。即ち音声素片選択部１１２は、上記のＣＶ単位の音声素片を音声素片メモリ１１３から順次読み出す。そして音声素片選択部１１２は、読み出した音声素片を音声素片接続部１１４に渡す。

【００７０】音声素片接続部１１４は、音声素片選択部

１１２から渡された音声素片を順次補間接続することにより、合成すべき音声の音韻パラメータ（特徴パラメータ）を生成する。

【００７１】以上のようにして、ピッチパターン生成処理部１０８によりピッチパターンが生成され、音声素片接続部１１４により音韻パラメータが生成されると、音声合成部１０２内の合成フィルタ部１１５が起動される。合成フィルタ処理部１１５は、無声区間ではホワイトノイズを、有声区間ではインパルスを駆動音源として、音韻パラメータであるケプストラム係数を直接フィルタ係数とするＬＭＡフィルタにより音声を出力する。

【００７２】以上、本発明の実施形態について説明したが、本発明は前記実施形態に限定されるものではない。例えば前記実施形態では、音声の特徴パラメータとしてケプストラムを使用しているが、ＬＰＣやＰＡＲＣＯＲ、フォルマントなど他のパラメータであっても、本発明は適用可能であり同様な効果が得られる。

【００７３】また、前記実施形態では、特徴パラメータを用いた分析合成型の方式の音声合成装置に実施した場合について説明したが、波形編集型やフォルマント合成型の音声合成装置であっても本発明は適用可能であり、やはり同様な効果が得られる。

【００７４】音韻継続時間長の制御に関しても、上述のような等時性を利用した方法でなくともよく、例えば統計的な手法を利用した場合でも本発明は適用可能である。更に、ユニットパターンの単位はアクセント句単位であることに限定されず、複数のアクセント句を含む、より長い単位でもよい。

【００７５】また、前記実施形態では、ユニットパターンをその形状と形状のピッチバイアス値に分離して扱い、テキスト解析部１０１でのテキスト解析結果に基づいて、ユニットパターン形状選択規則（第１の規則）に従いユニットパターン形状記憶部１０９からユニットパターン形状を選択すると共に、ユニットパターン形状のピッチバイアス量推定規則（第２の規則）に従いピッチバイアス量（ピッチバイアス値）を推定して、その選択したユニットパターンと推定したピッチバイアス量とからユニットパターンを生成する場合について説明したが、これに限るものではない。例えば、人間が発声した音声を分析して得られるピッチパターンを所定の関数に基づいたモデル（臨界制動２次線形系の応答関数に基づいたピッチ生成モデル、スプライン関数に基づいたモデルなど）によって近似し、近似に用いたモデルパラメータ、もしくは当該モデルパラメータより生成される近似されたピッチパターンから、所定の単位（例えばアクセント句単位）で切り出したユニットパターンをそのまま、もしくはクラスタリングによって得られる代表ユニットパターンの形で複数記憶しておき、その中からテキスト解析部１０１でのテキスト解析結果に基づいて所定の単位でユニットパターンを選択するようにしても構わない。

【００７６】また、前記実施形態では、音声合成装置が、パーソナルコンピュータ等の情報処理装置上で専用のソフトウェア（文音声変換ソフトウェア）を実行することにより実現されるものとして説明したが、専用のハードウェア装置により実現されるものであっても構わない。

【００７７】要するに本発明はその要旨に逸脱しない範囲で種々変形して実施することができる。

【００７８】

【発明の効果】以上詳述したように本発明によれば、少数のユニットパターンを用意するだけで、有声音韻と無声音韻の並びによらずに、全ての音韻の並びに適用でき、様々な文を対象とする音声合成が、ユニットパターンの無声区間と有声区間の境界におけるピッチの乱れに起因した音声合成時の揺れやピッチ制御の悪影響を生じることなく行える。

【００７９】また本発明によれば、少数のユニットパターンでありながら、ユニットパターンの形状とバイアス値を分離することによって、様々な合成内容に対応可能となる。

【図面の簡単な説明】

【図１】本発明の一実施形態に係る音声合成装置の概略構成を示すブロック図。

【図２】人間が発声した音声から抽出されるピッチパターンと、そのうちの有声区間と無声区間の境界付近に存在する細かなピッチ変化の例と、このピッチ変化を取り除くための近似されたピッチパターンの例と、この近似されたピッチパターンから切り出されたユニットパターンの例とを示す図。

【図３】ユニットパターン形状を選択しようとしているアクセント句のモーラ数以上で最も近いモーラ数のユニットパターン用の形状選択規則のうち、アクセント型が最も近いユニットパターン用の形状選択規則を用いて選択したユニットパターンと、そのユニットパターンに対してアクセント型が一致するように加えられた変形の例を示す図。

【図４】ユニットパターン形状候補の各組み合わせの評価手法を説明するための図。

【図５】従来の音声合成装置のブロック構成図。

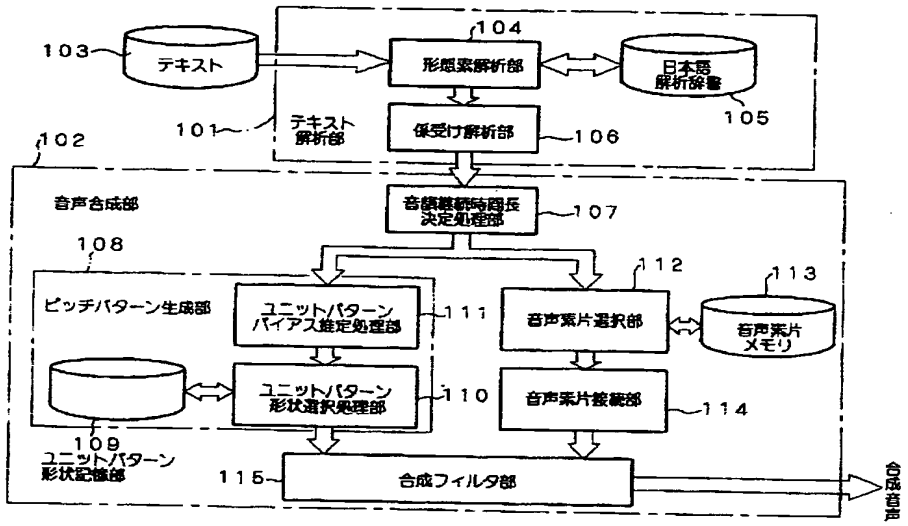
【図６】ピッチパターンを説明するための図。

【符号の説明】

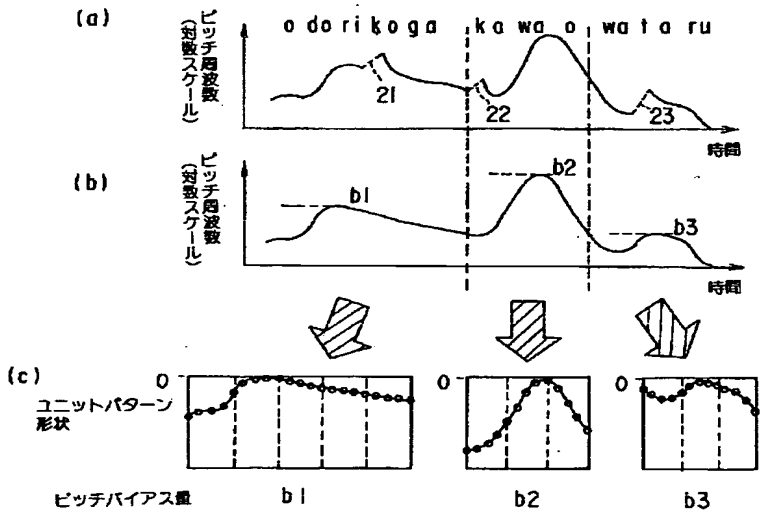
１０１…テキスト解析部
１０２…音声合成部
１０３…テキストファイル
１０４…形態素解析部
１０５…日本語解析辞書
１０６…係受け解析部
１０７…音韻継続時間長決定処理部
１０８…ピッチパターン生成処理部（ピッチパターン生

成手段）
１０９…ユニットパターン形状記憶部（ユニットパターン記憶手段、ユニットパターン形状記憶手段、記憶手段）
１１０…ユニットパターン形状選択処理部（ユニットパターン形状選択手段）
１１１…ユニットパターンバイアス推定処理部（ピッチバイアス値推定手段）
１１２…音声素片選択部
１１３…音声素片メモリ
１１４…音声素片接続部
１１５…合成フィルタ部

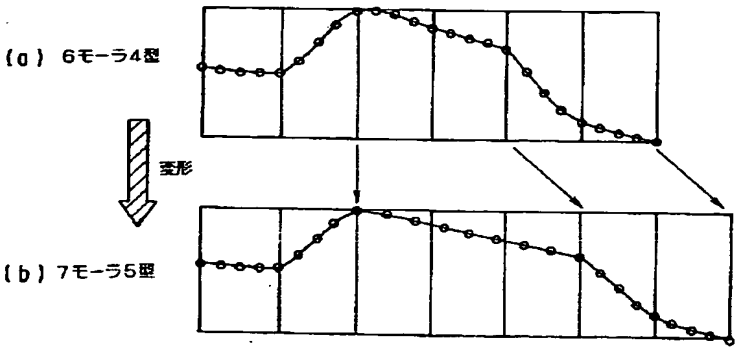
【図１】



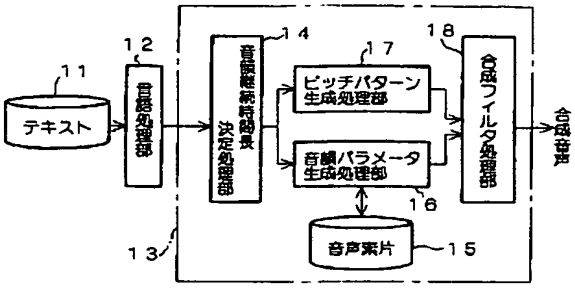
【図２】



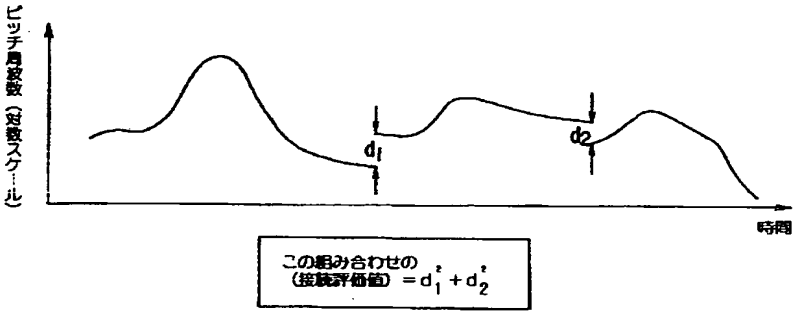
【図３】


(a) 6モーラ4型
変形
(b) 7モーラ5型

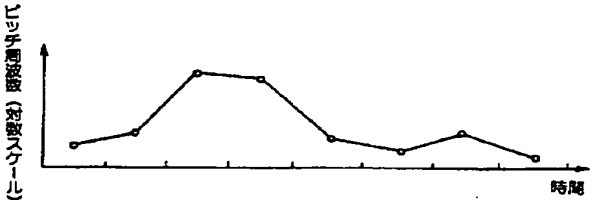
【図５】


テキスト
11
12
言語処理部
音韻継続時間長決定処理部
14
ピッチパターン生成処理部
17
音韻パラメータ生成処理部
16
18
合成フィルタ処理部
合成音声
音声素片
15
13

【図４】


ピッチ周波数（対数スケール）
d1
d2
時間
この組み合わせの（接続評価値）＝d1²＋d2²

【図６】


ピッチ周波数（対数スケール）
時間